图书情报工作
第 64 卷 第 9 期 2020 年 5 月

UGC 挖掘中的在线医疗社区分面体系构建与实现*

■ 翟姗姗[1] 潘英增[1] 胡畔[1] 许鑫[2]
[1] 华中师范大学信息管理学院 武汉 430079 [2] 华东师范大学经济与管理学部信息管理系 上海 200241

摘 要：[目的/意义]在线医疗社区是公众查询健康信息的主要方式之一。针对当今 Web2.0 模式下在线医疗社区中分面体系普遍存在的分面维度低、体系层级浅、分面焦点词不合理等问题，提出一个网络健康分面类型框架，以期改进在线医疗社区的分面导航，提升健康信息服务质量。[方法/过程]从 UGC 角度出发，结合用户关注健康信息主题与网络健康信息质量评价形成 18 个类别的网络健康信息分面类型框架。以有问必答网 - 全部问题区的 UGC 健康信息为数据源，构建在线医疗社区分面体系原型。[结果/结论]所构建分面体系原型能够在一定程度上改善现今分面体系的不足，为构建 Web2.0 模式下在线医疗社区分面体系提供一种可行方案。

关键词：UGC 健康信息 医疗社区 分面体系
分类号：G250
DOI：10.13266/j.issn.0252 - 3116.2020.09.013

1 引言

健康信息是指与医疗、预防、疾病、保健、康复、养生、健康教育等内容相关的信息[1]。近年来，“互联网 + 医疗”随着信息技术与网络技术的快速发展而蓬勃兴起，截至 2017 年底，中国互联网医疗的用户规模数量已达到 2.53 亿，占据全国网民的 32.7%[2]。互联网由于其方便、自由、成本低廉、数据量大等特点使得公众能方便、快捷地通过网络获取健康信息，我国各大门户网站的健康频道、健康医疗 APP 和专业医疗社区已成为公众最主要的健康信息来源。然而，当前多数 Web2.0 模式的社区没有建立分面体系，仅小部分建立了分面导航，但分面维度低，其中有一些社区只是采用了标签的方式对信息内容进行简单的分类，远远不能为用户提供满意的效果，为了让用户更容易、方便、高效地获取健康信息，需要以健康信息为数据源，构建逐步细化用户需求的人性化分面体系。

基于此，笔者尝试提出一种基于 UGC 的分面体系构建方案，其主要工作分布于分面框架构建与各分面焦点词确定。在实证部分，以有问必答网站 - 全部问题区的数据帖子为样本，对其进行 UGC 同义词识别，提取分面类型，确定各分面焦点词，设计分面体系系统原型。

2 国内外相关研究

2.1 用户生成内容(UGC)

用户生成内容(User Generated Content，UGC)是在 Web2.0 时代背景下产生并快速发展的一种信息资源创作与组织模式[3]。学术界对其已有研究，詹丽华以用户的数据素养和用户行为为出发点，得出了数据素养和用户行为会共同影响用户关注的 UGC 领域[4]；赵宇翔等深入解析 UGC，在类型理论基础上提出了 UGC 概念分析框架，完善了 UGC 的研究思路[5]；金燕提出了一种基于情绪分析的 UGC 质量评判模型，该模型能及时识别低质量 UGC，有助于舆情监控、规范网络秩序[6]；王晰巍等对移动图书馆 UGC 进行情感分析，得出将情感分析相关理论和方法引入移动图书馆 UGC 研究，能够改善移动图书馆的信息服务质量[7]；万力勇等构建了教育类 UGC 质量满意度概念模型并进行了实证检验，得出教育类 UGC 的完整性、可用性、丰富性、规范性和有效性是影响用户满意度的关键因素[8]；金燕等在 UGC 基础上识别异常行为，形成了基于用户画像的 UGC 质量预判

* 本文系国家社会科学基金青年项目“社会网络中基于用户认知结构的知识标注研究”(项目编号：17CTQ024)研究成果之一。
作者简介：翟姗姗(ORCID：0000 - 0002 - 2787 - 0183)，副教授，E - mail：zhais@ mail. ccnu. edu. cn；潘英增(ORCID：0000 - 0002 - 4502 - 3949)，硕士研究生；胡畔(ORCID：0000 - 0002 - 3916 - 5457)，硕士研究生；许鑫(ORCID：0000 - 0001 - 7020 - 3135)，教授，博士生导师。
收稿日期：2019 - 11 - 18 修回日期：2019 - 12 - 24 本文起止页码：114 - 121 本文责任编辑：徐健

模型[9]。

研究表明,UGC 中蕴含巨大的信息价值,当今学术界对 UGC 的研究主要集中在 UGC 情感、短文本关键词抽取与短文本质量方面。在线医疗社区中,用户既是健康信息资源的消费者,又是健康信息资源的生产者,如何从这些非结构化的 UGC 中抽取其所蕴含的有价值信息是文本处理所致力解决的问题。

2.2 健康信息组织与导航方式研究

信息组织是将信息从无序变为有序、系统的过程[10],是构建分面体系关键的一个步骤,信息组织的质量会影响分面导航的建立效果。学术界对此已有相关研究,王娜认为标准建设、管理监督、技术支持与用户参构成了泛在网络信息组织机制的要素[11];侯冠华等得出数字图书馆导航结构影响老年读者的情感体验、感知的可用性和任务绩效,且当认知负荷与导航结构共同作用时,会对老年读者的情感体验、阅读绩效产生显著的影响[12];王若佳等采用日志挖掘的方法研究查询和点击行为两个角度,提出了要注重以用户为中心的网站设计模式[13];陈果等从 UGC 角度出发,实现概念关联,融合知识库,构建了以丁香园心血管论坛为对象的分面导航体系,并实现了相应的原型系统[14];胡潜等利用相关评测标准分析行业用户对信息资源聚合的影响,提出面向用户的行业信息资源聚合模型[15];张鑫等从在线健康信息搜寻任务的角度出发,得出用户在线健康信息查询可依据通用切面和属性特征两个维度进行分类,并构建了一个分面分类理论模型[16];邱明辉采用文献调查法和文献分析法,从系统、任务、用户等方面对信息查询系统分面导航的设计和评价进行分析综合,形成一个信息查询系统分面导航设计的知识体系,并提出系统设计的相关建议[17]。

综上所述,针对信息组织与导航方式的研究中,或聚焦在信息文本内容的分析,或关注于用户行为的研究,或是分析信息组织与导航方式的影响因素,或是对当前分面组织与导航方式进行优化和改进。但在健康信息领域中,围绕健康信息组织与导航方式进行的研究存在一定的不足,如健康信息组织方式的研究视角缺乏系统比较、导航方式策略稍显单一等,使得分面体系尚未得到有效构建和利用。笔者从 UGC 的角度出发,提出一种分面体系构建方案,并设计原型系统验证其可行性。

3 分面体系模型设计

3.1 研究思路

分面体系又叫分面查询、分面搜索[18],按照“分面－亚面－类目”的规则排列,方便用户缩小、扩大查询范围或改变查询方向,满足用户交互式和探索式的检索行为。构建在线医疗社区分面体系需解决两个关键问题:第一,分面框架的构建;第二,各分面焦点词的确定。基于此,笔者采取的策略如下:一方面,通过用户对健康信息的关注主题与网络健康信息质量评价提取分面类型基本框架;另一方面,使用 UGC 同义词识别建立网络词与主题词之间的概念关联,结合“CMesh 主题词表＋知识库＋电子病历”确定各分面焦点词。整体研究框架如图 1 所示:

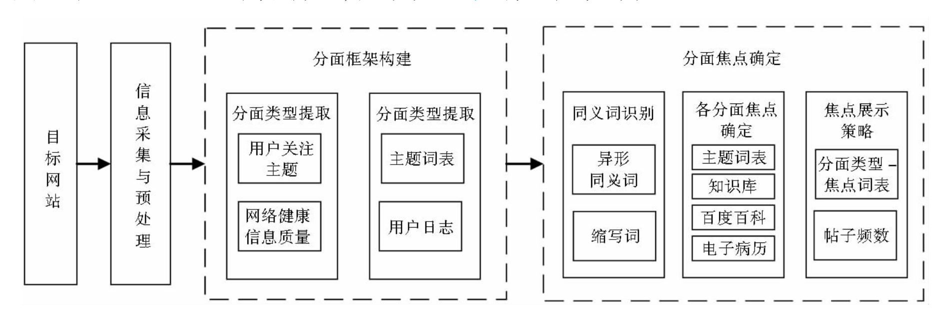


图 1　健康信息分面体系研究框架

3.2 分面框架构建

3.2.1 分面类型提取

分面类型是指描述一个分面的语词,Web2.0 模式下,用户可以在在线医疗社区中自由提问或回答,由此产生了大量的短文本健康信息,仅以传统知识库提取分面类型的书面性较强,难以关联非结构化的 UGC 信息,因而不适用于在线医疗社区。商丽丽等发现微信公众号中,用户对健康信息关注度较高的是健康风险、饮食、药物、身体活动和癌症主题类型[19]。笔者也调研了 12 家在线医疗社区发布的帖子发现,用户对健康问题的关注最为直接的是有没有患病、患的哪种病、患病严不严重、如何治疗该疾病、该用什么药、该用什么治疗手段、风险大不大、如何疾病控制、该领域权威专家等方面,因此,确认用户对健康信息关注的主题是提

取分面类型的关键。除此之外，社区是用户获取健康信息的主要来源，用户可以在在线医疗社区中搜寻自己所需的健康信息，因而网络健康信息的质量直接影响用户的健康搜寻体验和健康素养水平[19]。

本文中的分面类型结合了用户对健康信息的关注主题与网络健康信息质量评价两个方面，涉及了关于病、关于人、人的需求、帖子信息质量 4 个方面。关于病包括了疾病名、并发症、病发症状等，其中疾病是最基础的病情特征，但同一疾病可能有不同的症状，而且对诊疗、护理都有重要影响，部分疾病还会依据患者群体特征进行差异化诊疗和护理；关于人是指性别、年龄、病史、过敏源等，这是为了帮助清晰、准确地界定用户的病情；人的需求则包括诊断手段、治疗方式、日常护理、医生推荐、医院推荐等，其中，诊断手段包括化学检查手段、物理检查手段，治疗方式则有饮食治疗、药物治疗、手术治疗等，日常护理如血压、体重、血脂等，涵盖了诊治的具体手段、工具，以及诊断实施的主体。帖子信息质量是在参考国内关于网络健康信息质量评价指标的相关研究[20-21]的基础上，最终采用有无回复、权威性、有用性、及时性 4 个指标作为本文的网络健康信息质量评价指标。综上，形成了疾病、患者、需求、信息质量 4 个基本范畴 18 个类别的分面类型（见表 1），这 18 个类别构成了网络健康信息分面类型词表的基本框架。

表 1　分面医学主题词表的基本框架

基本范畴	类	具体内容
疾病	疾病名称	引起人体生理或心理不适的各类疾病名称的集合，存在亚类，如糖尿病分为Ⅰ型糖尿病、Ⅱ型糖尿病、妊娠糖尿病等亚类
	疾病症状	疾病发生时所表现的一系列生理或心理症状，如糖尿病的症状为多饮、多尿、多食和体重减轻等
	并发症	一种疾病在发展过程中引起的另一种疾病，如糖尿病易引发糖尿病肾病、冠心病等
患者	性别	患者的性别，分为男、女
	年龄	患者的年龄，一般以 5 年分区
	病史	患者的既往病史、现今病史
	家族遗传	某种疾病是否在患者的家族中发生过，如隔代遗传、父系遗传等
	过敏源	导致患者过敏的方式，如花粉过敏、药物过敏等
	特殊群体	某些群体具有明显区别于其他的特征，如孕妇、产妇
需求	检查方式	用于发现或确定某种疾病的手段，存在亚类，亚类可分为物理、化学、一般方式
	治疗方式	用于治疗疾病的手段，治疗方式多种，可分为饮食、药物、手术、心理等亚类
	日常护理	血糖、体重等指标控制、术后护理、运动习惯等
	医生推荐	向患者推荐相关领域的专家或权威医生，如患者选择糖尿病后，向其推荐糖尿病领域专家
	医院推荐	向患者推荐相关领域的权威医院，如患者选择糖尿病后，向其推荐糖尿病领域医院
信息质量	有无回复	指健康信息帖子是否有回复，这是在线医疗社区中最重要的一个评价指标
	权威性	健康信息回复者是否是该领域有声望的专家、帖子回复次数、帖子点击浏览次数
	有用性	用户咨询的健康信息是否被该用户采纳、回复内容的点赞次数
	及时性	健康信息更新频率，具体指用户咨询的健康信息的最后回复时间

3.2.2　分面展示策略

用户在在线医疗社区中习惯使用疾病、症状、药品、个人信息等信息进行提问，因此可使用该类分面类型可作为初始分面。而对于其他分面，为了让用户感受到分面体系的好用和易用，需要对其展示进行控制，而用户的行为习惯具有连续性，由此可以通过日志中记录的用户历史行为来推测用户未来行为[22]。首先，由于各分面中的焦点词来自于各在线医疗社区中用户提问的数据，若用户输入的主题词表达含义较为宽泛，则根据实际需要动态添加子分面；其次，分面是由语词组成的，而词与词之间具有层级、相关、等同关系，若分面之间具有层级关系，则将其按照上下级关系排列；最后，某一分面（通常是底层分面）下的子分面的焦点词数量较少，为减少焦点词的路径过深问题，则可以将该分面下子分面的焦点词上移，即在本层分面显示子分面的焦点词。

3.3　各分面焦点确定

3.3.1　UGC 同义词识别

UGC 同义词识别的目的主要是建立用户网络词与主题词表之间的概念关联，从而将词表中过于正式的主题词替换成贴近用户习惯的语词，UGC 同义词主要包括缩写词和异形同义词。缩写词又分为中文缩写和英文缩写，如“艾滋病”的英文缩写是“AIDS”；异形同义词主要是同一概念的不同表达，如“获得性免疫综合症”的别名是“艾滋病”，英文名称是“acquired immunodeficiency syndrome”。

异形同义词的识别一方面主要依靠现有的相关知识库辅助识别,不同类型的知识库往往会使用不同的概念表达同一语词,但在其概念同义词集中会较大概率出现重复表达的语词,基本思路如下:在异形概念词集合中有出现相同的主题词,就将其划为同义词。另一方面,由于一个语词在词语体系中与其他语词有着同位词、上位词、下位词的关系,利用这些知识库扩展异形同义词可以提高同义词的识别概率,如 CMesh 医学主题词表中的疾病名称体系,它标出了疾病之间的关系,如果两个语词之间的上位词或者下位词相同,就将其划为同义词。

在表达时使用缩写词是 UGC 文本的常见情况,由于用户用词的多样性使一个概念语词的缩写词形式多变,但这种情况并不是毫无规律可循,大部分的缩写词是从相对应的全称语词中截取字符形成的,例如从全称是"黏膜黑斑－息肉综合症"截取出"黑斑息肉病",需要注意的是截取的字符可以是连续性的也可以是非连续性的。基本思路如下:将缩写词在主题词表中查找,若没有命中,则将该缩写词拆分为单字,若在词典中被命中,则将该缩写词划为同义词。若某一缩写词命中了多个主题词,则将该缩写词与最短的命中主题词划为同义词。

3.3.2 各分面焦点确定

焦点词是分面体系的入口检索词,由于在线医疗社区中的用户多为一般用户,这些用户没有受过专门的训练、医学素养不高、表达口语化,因此,在选取焦点词时需贴近网络用户习惯,比如,对于同一概念,使用通俗化的语词代替书面语会更友好。依据 3.2 章节的分面类型框架确定焦点词并将其划分至所属的类别下。一方面,CMesh 医学主题词表、百度百科与知识库中结构化地存储了大量的健康信息,并模块化地展示了这些健康信息的相关属性,因此,利用 CMesh 医学主题词表、百度百科与知识库抽取焦点词并对其归类具有可行性。此外,由于 CMesh 医学主题词表、百度百科与知识库的更新周期较长,健康信息更新具有一定的滞后性,不能满足在线医疗社区中用户的需要,而电子病历可以实时对健康信息进行结构化的存储并模块化地展示健康信息的相关属性。因此,笔者采用"医学主题词表＋知识库＋电子病历"的方式确定焦点词并归类,具体策略如下:①抽取 CMesh 主题词表、百度百科与知识库中显示的模块化健康信息的相关属性,将其归类到分面类型框架中所属的类别下;②抽取电子病历中显示的模块化的健康信息的相关属性,将其归类到分面类型框架中所属的类别下;③将其合并,绘制成一个分面类型－焦点词表。流程图如图 2 所示:

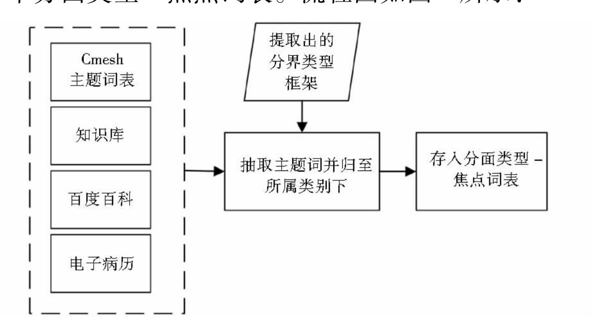


图 2 各分面焦点词确定策略

3.3.3 焦点词显示策略

在焦点词展示方面,为避免焦点词显示过多而造成的页面过载问题,需要对焦点词展示进行排序控制,若分面类型－焦点词表中已标记该分面下的焦点词在词表中具有层级的先后关系,则按照词表的顺序进行展示,否则,就按照焦点词的频次高低进行排序,以疾病分面举例,数据库中关于某一疾病(如老年痴呆、冠心病等)的帖子最多,则将该疾病(如老年痴呆、冠心病等)排在疾病分面中的最前面。此外,对于初次使用在线医疗社区的新用户,可以依据社区中多数用户的分面及焦点词的使用频次来进行展示控制。

4 分面体系原型实现——以有问必答网站为例

有问必答网(https://www.120ask.com/list/)是一个向公众提供健康问题咨询服务的网站,在中国医疗健康网站排行中位居第一,每日公众活跃度超过千万。该网站以科室分类,科室分类下又以全部问题、悬赏问题、以解决问题、待解决问题、零回答问题分区,这种导航方式反映了当前 Web2.0 健康信息网络社区中分面体系中存在的典型问题:分面维度低、体系层级浅、分面检索词不合理,不能满足用户渐进式、个性化、针对性的信息搜寻需求,笔者以全部问题区为例子,构建健康信息分面体系原型。

4.1 信息采集与预处理

使用 Java 中的 jsoup 技术采集了有问必答网－全部问题区中用户发布的帖子,采集字段内容包括帖子标题、帖子内容、发布时间、回复医生信息以及每一条回复内容,共采集了 9 433 个帖子,分别存储为标题、内容、链接 3 个属性。其中,内容字段包括医生姓名、医生职称、医生擅长领域、医生回复内容、医生回复时间。虽然采集的是医生与用户关于健康信息交流的网

站数据，但也会存在部分的无关帖子，如与健康信息无关的帖子，因此，需要对采集到的原始数据做预处理。使用以结巴分词工具与人工干预相结合的过滤策略，步骤如下：①使用结巴分词工具对帖子标题进行分词；②统计分词结果，筛选出医疗、预防、疾病、保健、康复、养生、健康教育相关关键词；③对原始数据进行分析，删除掉不包含这些相关关键词的帖子，剩余的 9 000 个帖子就形成了原型系统的数据源。

4.2 基于 UGC 的同义词抽取与入口词选取

通过网站调研，选择"有问必答网"中的疾病库、症状库、药品库、治疗库、手术库获取相关健康数据，又通过百度百科得到运动类别的数据，后在网络电子病历数据中识别相关类别主题词，对其归类并汇总形成最终的分面类型－焦点词表。具体数据如表 2 所示：

表 2 分面类型－焦点词表数量

分面类型	词数
疾病	2 410
症状	8 694
药物	3 380
检查	1 496
手术	10 610
食物	1 021
运动	165

UGC 同义词识别包括不同主题词之间的同义词识别，一方面，使用百度百科、疾病库等知识库为依据，结合 CMesh 医学主题词表中健康信息的主题词、受控词，进行异形同义词识别。以有问必答网中"癫痫"为例，百度百科中"癫痫"还被称为"羊角风""羊癫风"，疾病知识库中"癫痫"还根据病因分为了"反射性癫痫综合征""良性癫痫综合征""癫痫性脑病"，在 CMesh 医学主题词表中，"癫痫"的下位词包括了"发热性惊厥""局部性癫痫""全身性癫痫""创伤后癫痫""肌阵挛性癫痫""Landau-Kleffner 综合征""新生儿癫痫"等疾病。因此，对于疾病"癫痫"这一概念，需要将百度百科、疾病知识库、CMesh 医学主题词表中的"癫痫"相结合进行异形同义词识别。

另一方面，在 UGC 同义词识别时，需要将网络用户表达用语与主题词进行概念关联，如"腹痛"这一概念是被使用在主题词表中表达腹部疼痛，而网络社区中的用户习惯使用"肚子痛""肚子疼"来描述腹部疼痛，两者表达含义相同，在 UGC 同义词识别时，需要将"肚子疼""肚子痛""腹痛"划为同一概念。

而对于有问必答网中 UGC 的缩写词识别，则需要使用分词工具分词，以"内科"疾病为例，将预处理后的有问必答网－内科的问题帖子分成 5 个长文本，每个长文本包括 900－1 000 个帖子，将前文确定的分面类型－焦点词表导入作为词典辅助中文分词，对于词典中命中不了的新词作为候选缩写词，将该词切分成单字，逐一在词典中比对，从而实现 UGC 缩写词与主题词的概念关联，经过上述环节后得到的结果如表 3 所示：

表 3 有问必答网内科同义词识别结果（部分）

主题词	同义词
冠心病	冠状动脉粥样硬化性心脏病、缺血性心脏病、冠状动脉硬化、冠状动脉硬化性心脏病、冠状动脉心脏病
高血压	高血压症、血压高
老年痴呆	阿尔茨海默症、阿尔茨海默病
头晕	头昏、头胀、眩晕
恶心	反胃、胃部不适
腹痛	肚子疼、肚子痛

4.3 原型系统分面体系构建

原型系统的分面类型的选取依据 3.2 章节分面框架构建，以有问必答网－全部问题区的帖子数作为原型系统分面体系构建的数据来源，分面体系主要包括分面设计与焦点词的选择。在分面设计部分，一方面，分面体系设有检索框，若分面中不显示用户所需的健康信息，可在检索框中输入语词进行检索；另一方面，初始分面类型设置为疾病、症状、药品、高级选项 4 个分面，高级选项中设有性别、年龄、过敏源、是否回复、医生职位、回复状态、回答数、更新时间、医生推荐、医院推荐、其他 10 个子分面，后续的分面类型则根据用户具体点击动态呈现。在焦点词的选择部分，将经过预处理的帖子数据分词，进行 UGC 同义词识别后按照分面类型基本框架进行各主题词归类，将具有上下级关系的主题词按照层级进行排列，若提取出的焦点词表达含义过于宽泛，则根据需要将其设为新的亚目，一般来说，有些可以直接在基本框架下设置亚目，而有些则需要在第二级类目后才设置亚目。以老年痴呆为例的分面体系（部分）见表 4。

4.4 原型系统的检索结果展示与分析

为了比较原型系统与有问必答网分面体系效果，进行了分面检索结果对比分析，在对比过程中选择了"老年痴呆"常见疾病，其中图 3 与图 4 分别展示了有问必答网与原型系统中"老年痴呆"的分面检索结果截图，图 5 是原型系统"老年痴呆"后并发症"抑郁症"的检索结果截图，图 6 是原型系统"老年痴呆"后相关

表4　有问必答网老年痴呆分面体系(部分)

基本范畴	分面	焦点词
疾病	疾病名	
		老年痴呆
		……
	并发症	
		抑郁症
		脑萎缩
		出血
		水电解质紊乱
		泌尿道感染
		骨折
		肺部感染
		尿路感染
		褥疮
		吸入性肺炎
		胃肠道不适
	症状	
		记忆力衰退
		判断力障碍
		定向力障碍
		大小便失禁
		个性改变
		失语
		注意力不集中
		失认
		视空间能力下降
	……	

症状“记忆力衰退”的检索结果截图,图7是原型系统“老年痴呆”后相关症状“记忆力衰退”与“大小便失禁”的检索结果截图。

图3　有问必答网“老年痴呆”分面检索结果截图

图3是有问必答网“老年痴呆”的检索结果页面,所展现的分面方式反映的典型问题如下:①逻辑层次模糊,分类较为混乱。该分面体系按科室进行分面,分面下不仅显示内科、外科妇产科等科室分类标签,还出现了保健养生、运动瘦身、家居环境等不属于科室的分类标签。②分面单一,检索路径过深。该网站仅提供按科室分面,一级科室下又出现二级子科室,以“老年痴呆”为例,笔者经过了2层子分面才检索到该疾病。③焦点词与检索结果不匹配,帖子分类不合理。选择“老年痴呆”之后,检索结果出现不属与该疾病的帖子,“本人今年25岁,可是胸却像老人的胸”(见图3)。选择“老年痴呆”后的分面体系原型的检索页面截图(见图4),则向用户提供了并发症、相关症状、检查、药品、高级选项分面,点击分面中的“更多”可展开剩下的焦点词,“多选”是指可选择多个焦点词,各个分面的焦点词按帖子中该词出现的频次高低排序。

图4　原型系统“老年痴呆”分面检索结果截图



图5　原型系统点击“抑郁症”后的分面检索结果截图

图5是原型系统在选择疾病“老年痴呆”的前提下进行的关于并发症“抑郁症”的二次检索结果截图,由于老年痴呆与抑郁症是两种不同的疾病,点击“抑郁症”后,分面体系中的症状、诊断、药品分面类型的检索词都发生了相应的变化,检索得到的帖子标题既包含“老年痴呆”又包含了“抑郁症”,图6与图7分别是基

若找不到，请输入　搜索

并发症：抑郁症　脑萎缩　肺部感染　尿路感染　震颤　骨折　出血　水电解质紊乱　褥疮　更多>>　多选+

相关症状：记忆力衰退　判断力障碍　定向力障碍　个性改变　失语　注意力不集中　大小便失禁　失认　更多>>　多选+

检查：脑电图检查　CT检查　冠状位检查　脑磁共振检查　脑血流图　脑脊液检测　量表评估　血常规　更多>>　多选+

药品：多奈哌齐　银杏叶片　尼麦角林片　脑蛋白水解物　吡硫醇　尚拉西坦　石杉碱甲　去甲替林　更多>>　多选+

高级选项：性别-女　年龄　过敏源　医院推荐　医生推荐　有回复　回复状态　医生职位　其他　更多>>

标题	回答数	状态	更新时间
最近老是忘事情？会不会老年痴呆？	2	未采纳	5天前
老年痴呆症，话多，记忆力下降	3	未采纳	2018-9
记忆力下降会得老年痴呆吗	3	未采纳	8天前
老人胡说八道，记忆力下降，是不是老年痴呆？	4	未采纳	8天前
中年人记忆力衰退，是老年痴呆的前兆吗？	1	未采纳	2015-1
中年妇女记忆力严重衰退怎么办？	2	未采纳	3月前
我妈妈记忆力下降得很厉害了	2	已采纳	5月前
最近一段时间，记忆力严重减退	3	未采纳	2018-5

图 6　原型系统点击“记忆力衰退”后的分面检索结果截图

若找不到，请输入　搜索

并发症：抑郁症　脑萎缩　肺部感染　尿路感染　震颤　骨折　出血　水电解质紊乱　褥疮　更多>>　多选+

相关症状：记忆力衰退　判断力障碍　定向力障碍　个性改变　失语　注意力不集中　大小便失禁　失认　更多>>　多选+

检查：脑电图检查　CT检查　冠状位检查　脑磁共振检查　脑血流图　脑脊液检测　量表评估　血常规　更多>>　多选+

药品：多奈哌齐　银杏叶片　尼麦角林片　脑蛋白水解物　吡硫醇　尚拉西坦　石杉碱甲　去甲替林　更多>>　多选+

高级选项：性别-女　年龄　过敏源　医院推荐　医生推荐　有回复　回复状态　医生职位　其他　更多>>

标题	回答数	状态	更新时间
记忆力衰退，全身无力，大小便失禁，吐字不清	4	未采纳	8天前
记忆力减退，大小便控制不住，总是控制不住	7	未采纳	7天前
我妈妈有老年痴呆症，现在大小便失禁，给她	6	未采纳	7天前
老奶奶现在老年痴呆，大小便失禁	4	未采纳	8天前
老年痴呆症，话多，记忆力下降	3	未采纳	2018-9
我妈妈老年痴呆现在大小便失禁应该咋办有没有办法	5	未采纳	8天前
老人痴呆老人痴呆，大小便失禁，不能走路，	3	未采纳	5天前
老年痴呆患者大小便失禁脸色发青还能活多久	3	未采纳	2009-1
老年痴呆症大小便失禁治疗	3	未采纳	2016-5

图 7　原型系统点击“记忆力衰退”与“大小便失禁”后的分面检索结果截图

于“老年痴呆”的“记忆力衰退”与“性别 - 女”的组合检索结果截图和基于“老年痴呆”的“记忆力衰退”“大小便失禁”与“性别 - 女”的组合检索结果截图，由于检索条件比较类似，检索出了相同的帖子，即“老年痴呆症，话多，记忆力下降”。经过多个焦点词的组合检索之后，原型系统筛选掉了很多不相关的帖子，检索得到的帖子相关度变大，改善了健康信息的检索和结果，能够有效地帮助用户根据自身需要进行快捷检索。

5　结语

随着“互联网 +”医疗产业的产生和快速发展，Web2.0 模式下的在线医疗社区在未来会是用户搜寻健康信息的主要方式。另外，由于公众对自身健康越加重视，导致在线健康信息搜寻需求和行为增多，而现今在线医疗社区中的分面体系并未完全建立。基于这一背景，笔者提出了针对有问必答网 - 全部问题区的分面体系的构建方案，基于用户关注健康信息主题与网络健康信息质量评价的特点，进行 UGC 同义词的识别、分面框架的构建、各分面焦点词的确定以及展现策略，并构建了原型系统。实证表明，原型系统在很大程度上解决了原来在线医疗社区分面体系中分面维度低、体系层级浅、分面检索词不合理、资源覆盖率低的问题，提升了用户体验。

笔者研究并建立的在线医疗社区分面体系原型具有一定的通用性，分析和设计环节不受研究领域变化的影响，可以依据此研究思路应用于其他领域。但还存在一些不足，这也是后续研究中需要深入探讨的内容，主要包括健康文本主题抽取可视化、扩大原型系统的实际应用等。

参考文献：

[1] 张馨遥. 健康信息需求研究的内容与意义[J]. 医学与社会，2010，23(1)：51 - 53.

[2] 中国互联网发展报告(2018)[EB/OL]. [2019 - 08 - 06]. http://www.cac.gov.cn/2018-11/06/c_1123672145.htm.

[3] 肖强，朱庆华. 用户生成内容共享意愿的影响因素实证性研究[J]. 情报杂志，2012，31(4)：138 - 142.

[4] 詹丽华. UGC 用户行为成因分析——用户数据素养与用户行为情景的双重视角[J]. 情报理论与实践，2018，41(4)：28 - 32，37.

[5] 赵宇翔，范哲，朱庆华. 用户生成内容(UGC)概念解析及研究进展[J]. 中国图书馆学报，2012，38(5)：68 - 81.

[6] 金燕. 基于情绪分析的 UGC 质量评判模型[J]. 图书情报工作，2017，61(20)：131 - 139.

[7] 王晰巍，杨梦晴，韦雅楠，等. 基于情感分析的移动图书馆用户生成内容评价效果研究[J]. 图书情报工作，2018，62(18)：16 - 23.

[8] 万力勇，杜静，舒艾. 教育类 UGC 质量满意度影响因素实证研究——基于扩展的 ACSI 模型[J]. 中国电化教育，2019(3)：72 - 80.

[9] 金燕，孙佳佳. 基于用户画像的 UGC 质量预判模型[J]. 情报理论与实践，2019，42(10)：77 - 83.

[10] 娄策群，段尧清，张凯. 信息管理学基础第二版[M]. 北京：科学出版社，2009：131.

[11] 王娜. 泛在网络中的信息组织机制研究[J]. 现代情报，2018，38(5)：25 - 31，36.

[12] 侯冠华，董华，刘颖，等. 导航结构与认知负荷对老年读者数字图书馆用户体验影响的实证研究——以国家数字图书馆为例[J]. 图书情报工作，2018，62(13)：45 - 53.

[13] 王若佳，李培. 基于日志挖掘的用户健康信息检索行为研究[J]. 图书情报工作，2015，59(11)：111 - 118.

[14] 陈果，肖璐，孙建军. 面向网络社区的分面式导航体系构建——以丁香园心血管论坛为例[J]. 情报理论与实践，2017，40(10)：112 - 116.

[15] 胡潜，李静. 面向用户的行业信息资源聚合研究——以母婴健康行业用户知识社区为例[J]. 图书情报知识，2018(1)：87 -

94.
[16] 张鑫,王丹. 用户在线健康信息搜寻任务研究[J]. 情报资料工作,2017(6):74－83.
[17] 邱明辉. 信息查询系统的分面导航设计研究[J]. 现代情报,2018,38(10):78－84,120.
[18] 商丽丽,王涛. 基于用户信息行为的微信健康信息关注度研究[J]. 情报科学,2019,37(8):132－138.
[19] 邓胜利,赵海平. 用户视角下网络健康信息质量评价标准框架构建研究[J]. 图书情报工作,2017,61(21):30－39.
[20] 姜雯,许鑫,武高峰. 附加情感特征的在线问答社区信息质量自动化评价[J]. 图书情报工作,2015,59(4):100－105.
[21] 钱明辉,徐志轩,连漪. 在线健康咨询平台信息质量评价及其品牌化启示[J]. 情报资料工作,2018(3):57－63.
[22] 胡昌平,林鑫. 科技文献检索中基于主题词表分面化改造的分面构建[J]. 情报学报,2015,34(8):875－884.

作者贡献说明：
翟姗姗：提出研究思路与研究框架；
潘英增：数据采集与分析、撰写论文初稿；
胡畔：实证平台的原型开发；
许鑫：提出论文修改意见并修订最终版。

Study on the Construction and Implementation of Online Medical Community Faceted System in UGC Mining

Zhai Shanshan[1]　Pan Yingzeng[1]　Hu Pan[1]　Xu Xin[2]
[1] School of Information Management, Central China Normal University, Wuhan 430079
[2] Department of Information Science, Business School, East China Normal University, Shanghai 200241

Abstract: [**Purpose/significance**] The online medical community is one of the main ways for the public to query health information. Aiming at the problems of few faceted dimensions, simply system level, unreasonably search terms of the faceted system in the online medical community under the current Web2.0 model, this paper puts forward the network health faceted type framework, in order to improve the faceted navigation of the online medical community, and better its information services quality. [**Method/process**] This paper, from the perspective of UGC, obtained 18 categories of network health information faceted type framework, which combines with the user's attention to the health information topic and the evaluation of the network health information quality, and constructed a prototype of the healthy information faceted system that data sources comes from the QuanBuWenTi area of YouWenBiDa website. [**Result/conclusion**] The experimental result shows that the prototype of the faceted system constructed in this paper can effectively improves the shortcomings of the current faceted systems, and provides a feasible solution for building faceted system of the online medical community under the Web2.0 mode.

Keywords: UGC　healthy information　medical community　faceted system